Mina Fallahi, Fabian Brinkmann, Stefan Weinzierl

Simulation and analysis of measurement techniques for the fast acquisition of head-related transfer functions

Fallahi, Mina; Brinkmann, Fabian; Weinzierl, Stefan (2015): Simulation and analysis of measurement techniques for the fast acquisition of head-related transfer functions. In: Fortschritte der Akustik - DAGA 2015: 41. Jahrestagung für Akustik, 16. - 19. März 2015 in Nürnberg. Berlin: Deutsche Gesellschaft für Akustik e.V. pp. 1107–1110.

Simulation and analysis of measurement techniques for the fast acquisition of head-related transfer functions

Mina Fallahi[1],Fabian Brinkmann[1], Stefan Weinzierl[1]

[1] *Audio Communication Group TU Berlin, Einsteinufer 17c D-10587 Berlin, Germany*
E-Mail: mina.fallahi@mailbox.tu-berlin.de, {fabian.brinkmann, stefan.weinzierl}@tu-berlin.de

Introduction

Head-related transfer functions (HRTFs) describe the free field sound propagation between a sound source and the listener's ears and include all the cues which are evaluated forspatial hearing. If non-individual HRTFs are used to reproduce binaural signals at a listener's ears, localization and coloration errors occur [1], due to the non-individual head, pinna and torso morphology. The measurement of individual HRTFs, however, is a challenge both with respect to the mechanical measurement setup, and an efficient signal (post) processing for the acquisition of a multitude of impulse responses for different angles of incidence. In the current study, two techniques aiming at a reduction of the measurement time without loss of quality were compared: The Optimized Multiple Exponential Sweep method (O-MESM) [2] and Normalized Least Mean Square (NLMS) adaptive filtering [3]. Because a systematic variation of measurement conditions such as SNR, THD, rotation speed of the subject, background noise level, and number of sound sources is hardly feasible in an actual measurement setup, we simulated the measurementsnumerically. Results suggest that a high resolution HRTF dataset can be measured within one minute using NLMS and within 10-13 minutes using O-MESM. In a second study, we verified the outcome of the simulation in an experimental setup [4].

Fast HRTF acquisition

Over the last years, two methods for an accelerated measurement of multi-channel audio systems were proposed: (Optimized) MESM and NLMS.

In principle, O-MESM [2] is a conventional FFT based sweep measurement [5]. For acceleration,excitations of subsequent channels are interleaved in time with a delay of τ_W, thus taking advantage of the temporal structure of impulse responses related toweakly nonlinear systems after excitation with an exponential sweep. O-MESM further improves the measurement speed compared to MESM [6] for the case that only a small percentage of the measured impulse response (with the length τ_{ir}) is of interest, which is the case with HRTFs having the length $\tau_{DUT} \ll \tau_{ir}$. This part of the impulse response can be protected from interference with nonlinear impulse responses of order k (with thelength $\tau_{ir,k}$) of neighboring channels, if the following constraint for τ_W is fulfilled [2]:

$$\tau_{DUT} + \tau_{sp} \leq \left(-\frac{\ln k}{r_s} \bmod \tau_W\right) \leq \tau_W - \tau_{sp} - \tau_{ir,k} \quad (1)$$

Here, τ_{sp} is an additional safety zone before and after the relevant part of the impulse response, τ_W the time delay between two subsequent excitations, and r_s the sweep rate.

NLMS adaptive filter system identification was proposed to acquire HRTFswhile continuously rotating a subject above the vertical axis, and thus obtaining a quasi-continuous azimuthal resolutionduring rotation [3]. The resolution in elevation is commonly given by a fixed number of loudspeakers mounted on a (semi)-circular arc.In this case, the estimated head-related impulse response (HRIR) $\hat{\boldsymbol{h}}_\phi^{l/r}(n+1)$ for the left/right ear, and elevation ϕ is iteratively calculated:

$$\hat{\boldsymbol{h}}_\phi^{l/r}(n+1) = \hat{\boldsymbol{h}}_\phi^{l/r}(n) + \mu \frac{e^{l,r}(n)\boldsymbol{x}_\phi^T(n)}{\sum_\phi \|\boldsymbol{x}_\phi(n)\|^2} \quad (2)$$

$$e^{l,r}(n) = y^{l/r}(n) - \underbrace{\sum_\phi \hat{\boldsymbol{h}}_\phi^{l/r}(n)\boldsymbol{x}_\phi(n)}_{=\hat{y}^{l/r}(n)} \quad (3)$$

The iteration minimizes the LMS error $e^{l,r}(n)$ between the current sample n of the binaural signal $y^{l/r}(n)$ recorded at the ear-canals of a subject and itsestimated counterpart $\hat{y}^{l/r}(n)$. The estimation is given by the sum in eq. (3), which represents one sample of a time-domain convolution process using the currently estimated HRIR and the past N samples of the excitation signal $\boldsymbol{x}_\phi(n)$. The adaption speedcan be optimized by exciting the system with a perfect sweep[7],and is controlled by the so called step size $0 \leq \mu \leq 1$ .Please note that bold symbols in eq. (2-3) represent vectors of length N (length of the HRIR), and that n can be translated into the azimuth θ_n.

Modeling HRTF measurements

Modeling HRTF measurementsequals modeling the pressure signals at the ears of a subject that result from excitation of the loudspeakers and rotation of the subject, followed by system identification with O-MESM orNLMS. Due to the rotation, the pressure signals are given by a non-stationary combination[8] of the excitation signal and the current HRIR[9]

$$y^{l/r}(n) = \left[\sum\nolimits_\phi \boldsymbol{x}_\phi^T(n)\boldsymbol{h}_\phi^{l/r}(n)\right] + n^{l/r}(n), \quad (4)$$

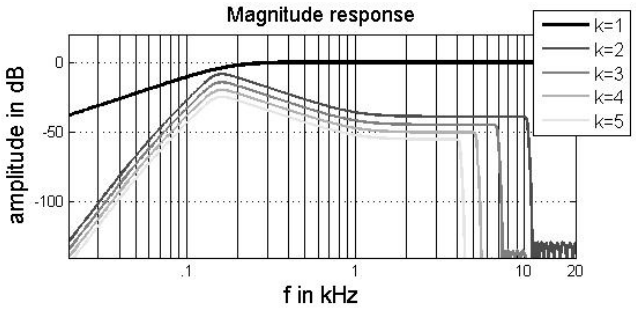

Figure 1: Linear and harmonic transfer functions of the modeled loudspeaker (THD = 3%, k=5).

where $n^{l/r}(n)$ is the environmental noise.Consequently, the applied model consisted of three parts: (a) The loudspeakers used for playback of the excitation signal, (b) the HRIRs representing the subject, the positions of the loudspeakers, and the acoustic transmission from the loudspeaker to the subjects' ears, and (c) the environmental noise.

The loudspeaker was considered to be the only source of non-linearity, and was thus modeled by a linear transfer function as well as a set of harmonic transfer functions. The linear part was approximated with a second order Butterworth high pass filter with a cutoff frequency at 180 Hz. To model the harmonic transfer functions, a combination of asecond order low-shelf filter(shelf frequency 1 kHz, gain 50 dB)and an eight order Butterworth high pass (cut off 150 Hz)were applied. The cutoff frequencies and gains of the filters as well as the level of the harmonic transfer functions relative to each other were obtained by visually fitting the loudspeaker model to measurements of typical 2" closed box loudspeakers from [4](cf. Fig. 1).The level of the harmonic transfer functions relative to the linear part was set according to the desired THD. The loudspeaker was considered weakly nonlinear with harmonic transfer function up to order $k = 5$ and total harmonic distortion (THD) up to 3% (≈ -30 dB). The effect of the loudspeaker model on arbitrary input signals was then calculated using the generalized Hammerstein model of non-linearity [10].

A high resolution and nearly full spherical HRIR dataset of the FABIAN head and torso simulator covering elevations between -64° and 90°[11]was used as the representation of subject inside the virtual measurement system.The dataset will be referred to as *reference HRIRs* in the following.In order to take into accountthe time variance of the system due to the continuous rotation of the subject, an HRIR corresponding to the current position $\theta_n$was interpolated for each sample *n*by inverse distance weighting[12]. This HRIR was then used to calculate the current sample of the microphone signal using non-stationary combination [8](cf. eq. 4)and the excitation signal including nonlinearities added in the previous model stage.

The environmental noise was modeled by a normally distributed random sequence witha spectral shapeaccording to a noise floor measured in [11] (first order low shelve, 1 kHz cut-off frequency, 35 dB gain).Theresulting noise was then applied to the modeled microphone signals according to eq. (4).The absolute level of the noise was set to obtain a desired peak-to-tail SNR (signal-to-noise ratio) if measuring a single HRIRwith an 16th order exponential sweep between 20and 22050 Hz. The single HRIR measurement was modeled by convolution of the sweep signal andthe left ear HRIR from the reference dataset ($\theta = 90°; \phi = 0°$), followed by addition of the noise and spectral deconvolution.

Although O-MESM, as originally proposed [2], is intended for a discrete azimuth measurement, it can also beused with a continuously rotating subject. In this case however, interpolation is needed to obtain HRIRs at the desired positions, given by the reference dataset.Because the subject rotates during the measurement, two constraints were applied on the rotation speed: (a)the angular rotation of the subject during the playback of a single sweep was limited to 1°,which approximately equals the localization blur for sources in the horizontal plane[13]. This constraint was set to achievedistinct localization performance if using the obtained HRIRs in virtual acoustic environments. (b)the angular rotation of the subject between two consequent excitation of the same loudspeaker was limited to 2°, aiming at a sufficient azimuthal resolution in orderto keepinterpolation artifacts below the threshold of perception[14]. These two constraints resulted in revolution times between 10 and 13 minutes for O-MESM depending on THD and SNR.The corresponding values for r_s and $\tau_W$were calculated using a slightly modified version of the ITA-Toolbox optimize method [15]. Because high frequencies are subject to greater spatial fluctuations due to the pinna fine structure, the azimuth used for interpolation was defined by the point where the excitation signal met6.92 kHz (geometric mean between 3 and 16 kHz).

Since HRIRs were obtained for every sample if using NLMS, there are no constraints on the rotation speed in this case and it was thus varied over a wide range. Moreover, the SNR and THD were varied in practically relevant ranges.The length of the HRIR was set to 3.5–4 ms.Table 1 summarizes the parameter set used for the simulations in this study. A complete variation of the parameters led to 270 simulated HRTF measurements in total.

Evaluation criteria

The output of the simulationsconsisted of HRIR datasets with source positions being a subset of the positions found in the reference. The evaluation of the results was based on the comparison to the reference HRTF dataset which was also used to model the microphone signals. The aim was to find out, to what extent the results of the modeled measurements differ from a traditional system-by-system HRTF measurement regarding interaural time and level differences (ITD, ILD) as well as spectral differences ITDs were assessed by means of differences in the broad-band time of arrival (TOA) between corresponding HRIRs of the left and right ear. The TOA was estimated using the ten times upsampled HRIRs and a simple threshold (-6 dB) with respect to the maximum value of each HRIR. The broad-band ILD was calculated as difference in logarithmic root mean square levelin dB between left and right ear HRIRs. Spectral differences were evaluated in $N_{fc} = 37$ equivalentrectangular bandwidth (ERB) auditory filters [16] as implemented in the auditory toolbox [17]

Table 1: Parameters used for simulation

	O-MESM	NLMS
Excitation	Exp. Sweep (20-22050Hz)	Perfect Sweep (0-22050Hz)
HRIR length	$\tau_{DUT} = 0.004$ s (176 samples)	$N = 0.0035$ s (156 samples)
Rotation time	10 to 13 min.	1, 5, 15 min.
SNR	90, 60, 40 dB	∞, 90, 60 dB
Method specific	$\tau_{ir} = 0.01$ s $\tau_{sp} = 0.001$ s $3 \leq r_s \leq 6$	μ : 0.25, 0.5, 1
Common	Number of channels: 10, 20, 39 THD: 0%, 1%, 3% Order of nonlinearities: 5	

$$E(f_c) = 10\log_{10}\frac{\int C(f,f_c)\,|HRTF_{sim.}|^2 df}{\int C(f,f_c)\,|HRTF_{ref.}|^2 df} \qquad (5)$$

Here, $C(f,f_c)$ is an auditory filter at center frequency $180\text{ Hz} \leq f_c \leq 20\text{ kHz}$. The results for the left and the right ear were then added and averaged over auditory filters

$$ERBerror = \frac{1}{N_{fc}}\sum_{fc}|E(f_c)_{left}| + |E(f_c)_{right}| \qquad (6)$$

A similar error measure prooved to be a good predictor for the audibility of interpolation artifacts in HRIRs [13]. Based on earlier perceptual evaluations [18], ERBerrors in the range of 0.5 to 1 dB were considered tolerable but slightly audible. ITD and ILD errors were considered tolerable, if they were in the order of the threshold of perception given by11 µs, and0.6 dB, respectively[14].

Simulation results and discussion

Simulation results for O-MESM and NLMSshowed almost no effect of the THD on the error measures (max. deviations: ERB 0.25 dB; ITD 2.26 µs; ILD 0.05 dB).THD was thus discarded from the discussion and instead results for the highest THD of 3% are shown. Moreover, O-MESMwas robusttowards variation in SNR (max. deviation: ERB 0.06 dB; ILD 0.03 dB and no deviation in ITD), which in this case was also excluded fromthe discussion and only results for 40 dB SNR are given in Fig. 2. Favored by the restrictions on the rotation speed posted above, ITD, ILD, and ERB errors are within the tolerable range in any case.

Please note that we chose to display the 95% percentile value, which we believe is a conservative estimator for the overall error.According to [2], O-MESM showed a robust behavior against the presence of environmental noise and nonlinear distortions, and is comparable

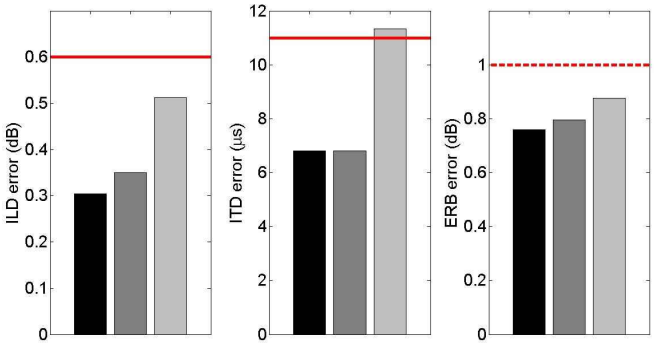

Figure 2: O-MESM performance: Comparison to the thresholds (red lines): 95%-percentile vales in the case of 40dB peak SNR and 3% THD for ILD error (left), ITD error (middle) and ERB error (right), for 10 (black), 20 (dark grey), and 39 channels (light grey).

to a sequential exponential sweep measurement. Differences between measurements with 10, 20, and 39 channels were believed to solely originate from differences in the rotation time:Slower rotation times result in larger angular movements of the subject between two consequent excitation of the same loudspeaker.Consequently, HRIRs are available in a coarser grid and larger errors occurwhen interpolating onto the grid defined by the reference dataset.

The results for NLMS adaptive filtering are depicted in Fig. 3.They show a general increase of errors with increasing number of loudspeaker channels as well as a general decrease with increasing rotation time and step size. Results for 15 minute rotation were omitted because they were almost identical to those obtained for a five minute rotation (max. improvement:ERB 0.2 dB; ITD 2.26 µs; ILD 0.25 dB). In addition, due to very small differences between the results of infinite and 90 dB SNR (max. deviation: ERB 0.01 dB; ITD 2.26 µs; ILD 0.01 dB) the results of thecase of noiseless environment are not shown. As can be seen fromFig. 3,even for a rotation time of only one minute, errors fall within the acceptable range if the measurement environment is relatively noiseless (SNR ≥ 90 dB) and the step size is set somewhere between 0.5 and 1. Noisier environments (SNR ≈ 60 dB) demand slower rotation of the subject. However, in this case the ERB error might be unacceptably large in any case, whereas errors in ITD and ILD are within the tolerable limit.

Conclusion and outlook

A system for the fast and high resolution measurement of individual HRTFs was simulated and evaluated using NLMS and O-MESM for system identification. A constant rotation of the subject during the measurements with up to 39 channels was simulated using a high resolution HRTF dataset and non-stationary combination.

Both system identification methods showed robustness towards changes in THD, however, only O-MESM was also robust against environmental noise. Nevertheless, the NLMSalgorithm offered the better option as long as the SNR wassufficiently high (SNR ≥ 90 dB). In this case, 5716 HRTFs for 39 elevations between -64° and 88° could be measured within one minute, whereas O-MESM required measurement durations between 10 and 13 minutes, which in turn should be favored in relatively noisy environments. The increased

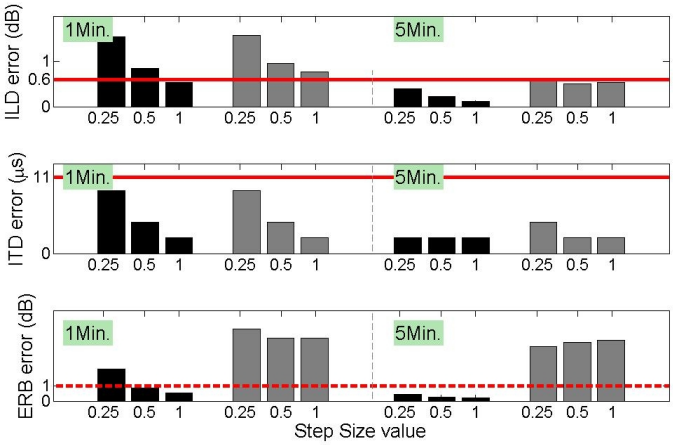

Figure 3: NLMS performance for 39 loudspeaker channels: Comparison to thresholds (red lines): 95%-percentile values for ILD error (top), ITD error (middle), ERB error (bottom), with 1 (left) and 5 minutes (right) rotation time, 90 (black), and 60dB SNR(grey) and for step sizesof 0.25, 0.5, and 1.

rotation times for O-MESM were related to a constraint on the rotation speed.This constraint is due to the need for interpolating the O-MESM results to the desired positions, whereas NLMS inherently offers data in a quasi-continuous azimuthal resolution. Interpolation errors could however be reduced, if applying interpolation in the frequency domain and using correct azimuths θ_n for each frequency bin, corresponding to the subject's position during excitation. Future studies could also investigate the dependency of the ERB error on the SNR in the range of $60 \leq \mathrm{SNR} \leq 90$ dB, for NLMS, as well as improvements in the adaptive system identification method [19-20].

Acknowledgement

The work is part of the Simulation and Evaluation of Acoustical Environments (SEACEN) project funded by the German Research Foundation (DFG WE 4057/3-2).

References

[1] Møller, H., Sørensen, M.F., Jensen, C.B.,Hammershøi, D.: Binaural technique. Do we need individual recordings?*J. Audio Eng. Soc.* 44**(6)** (1996), 451-469.

[2] Dietrich, P., Masiero, B.,Vorländer, M.: On the optimization of the multiple exponential sweep method. *J. Audio Eng. Soc.* 61**(3)** (2013), 113-124.

[3] Enzner, G.: 3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2009*, New Paltz, NY.

[4] Fuß, A., Brinkmann, F. and Weinzierl, S.: A full-spherical multi-channel measurement system for the fast acquisition of head-related transfer functions. *Fortschritte der Akustik - DAGA 2015*, Nürnberg, Germany.

[5] Farina, A.:Simultaneous Measurement of Impulse Response and Distortion with Swept-sine technique. *In Proc. 108th AES Convention*, Paris, France, 2000.

[6] Majdak, P., Balazs, P.,Laback, B.: Multiple Exponential Sweep Method for fast measurement of head-related transfer functions. *J. Audio Eng. Soc.* 55**(7-8)** (2007) 623-637.

[7] Telle, A., Antweiler, C., Vary, P.: Der perfekte Sweep-ein neues Anregungssignal zur adaptiven Systemidentifikation zeitvarianter akustische Systeme. *Fortschritte der Akustik - DAGA 2010*, Berlin, Germany, 341-342.

[8] Margrave, G.F.: Theory of nonstationary linear filtering in the fourier domain with application to time-variant filtering. *Geophysics* 63**(1)**(1998), 244-259.

[9]Enzner, G., Antweiler C.,Spors, S.: Trends in acquisitionof individual head-relatedtransferfunctions. In: *Technology ofbinaurallistening*. Blauert, J. (Ed.). Springer, Berlin. 2013 pp. 57-92.

[10] Novák, A., Simon, L.: Nonlinear system identification using exponential sweep-sine signal. *IEEE Transactions on Instrumentation and Measurement,* 59**(8)** (2010), 2220-2229.

[11] Brinkmann, F., Lindau, A.,Weinzierl, S., Geissler, G., van de Par, S. High resolution head-related transfer function data base including different orientations of head above the torso. *International Conference on Acoustics AIA-DAGA 2013*, Merano, 596-599.

[12] Hartung K., Braasch, J. andSerbing, S.J.: Comparisonof different methodsfortheinterpolationofhead-relatedtransferfunctions. *16th Int. AES Conf.*(1999),319-329.

[13] Blauert, J.:*Spatial hearing: the psychophysics of human sound localization.* MIT Press, Massachusetts, 1997.

[14] Minnaar, P., Plogsties, J., Christensen, F.: Directional resolution of head-related transfer functions required in binaural synthesis. *J. Audio Eng. Soc.* 53**(10)** (2005), 919-929.

[15] ITA-Toolbox, for MATLAB, Institut of Rechnical Acoustics, RWTH Aachen, URL: `http://www.ita-toolbox.org/`

[16] Moore, B.C.J.: Frequency analysis and masking. In: *Hearing Handbook of Perception and Cognition* Moore, B.C.J. (Ed.). Academic Press, San Diego, 1995, 161-205

[17] Slaney, M.: *Auditory toolbox - Version 2*.Tech. Report#1998-010,Interval Research Corporation, 1998.

[18] Brinkmann, F., Roden, R., Lindau, A.,Weinzierl, S.: Audibility of head-above-torso orientation in head-related transfer functions. In *Forum Acusticum*, Krakau, Poland, 2014.

[19] Antweiler, C.,Kühl, S.,Sauert, B., Vary, P.: System identification with perfect sequence excitation - Efficient NLMS vs. inverse cyclic convolution. *ITG-Fachbericht 252: Speech Communication* (2014), VDE Verlag GmbH, Berlin, Offenbach, Germany.

[20] Hahn, N., Spors, S.: Identification of dynamic acoustic systems by orthogonal expansion of time-variant impulse responses.*IEEE 6th International Symposium on Control, Communications, and Signal Processing (ISCCSP 2014)*, Athens, Greece, 161-164.